THE ANALYST.

VOL. X. JANUARY, 1883. No. 1.

ON AN UNSYMMETRICAL PROBABILITY CURVE.

BY E. L. DE FOREST.

[Continued from page 168, Vol. IX.]

We will now illustrate the applicability of the gamma curve to represent series which are not expansions of any known polynomial, but are simply the results of repeated observation of some phenomenon or occurrence, in which there is a manifest inequality in the distribution of the errors or deviations on either side of the mean. Take for example the observations given by Quetelet in his *Letters* already cited, of the amplitude of diurnal variation of temperature (centigrade) at Brussels in the month of January, as observed for a period of 10 years, from 1833 to 1842. Column (1) of

TABLE II.

(1) Amp.	(2) Days	(3) g	(4)	(5)	(6) x	(7) gx^2	(8) gx^3	(9) y	(10) $g-y$
0° to 1°	0	.000	0.5	.000	—4.7	.000	— .000	.000	—.000
1 2	8	.026	1.5	.039	—3.7	.356	—1.317	.028	—.002
2 3	31	.100	2.5	.250	—2.7	.729	—1.968	.110	—.010
3 4	61	.197	3.5	.689	—1.7	.569	— .967	.183	+.014
4 5	68	.220	4.5	.990	—0.7	.108	— .075	.200	+.020
5 6	50	.162	5.5	.891	0.3	.015	.004	.171	—.009
6 7	32	.104	6.5	.676	1.3	.176	.229	.124	—.020
7 8	22	.071	7.5	.533	2.3	.376	.864	.081	—.010
8 9	20	.065	8.5	.552	3.3	.708	2.336	.048	+.017
9 10	8	.026	9.5	.247	4.3	.481	2.067	.027	—.001
10 11	4	.013	10.5	.137	5.3	.365	1.935	.014	—.001
11 12	3	.010	11.5	.115	6.3	.397	2.500	.007	+.003
12 13	1	.003	12.5	.037	7.3	.160	1.167	.004	—.001
13 14	1	.003	13.5	.041	8.3	.207	1.715	.002	+.001
14 15	0	.000	14.5	.000	9.3	.000	.000	.001	—.001
	309	1.000		5.197		4.647	8.490	1.000	

our Table II. shows the various amplitudes, grouped by differences of 1°, the smallest one observed being between 1° and 2°, while the largest was between 13° and 14°. Column (2) shows the number of days on which each of the different amplitudes occurred. The whole number of days of observation was 309. The number of days to each amplitude being divided by 309, the quot's g are set in column (3). These may be regarded as the approximate probabilities that, in any future observation, the corresponding amplitudes will occur. The sum of all these probabilities is unity.

In any probability curve, either (52) or (57), the expression for y contains dx as a factor, so that we may if we please, regard y not as an ordinate, but as the differential Ydx of the area of a curve, that is, the area included between any two consecutive ordinates Y, whose abscissas are $x - \frac{1}{2}dx$ and $x + \frac{1}{2}dx$. Under this view, the total area of the probability curve is unity. The numbers g in column (3) approximately represent finite sections of this area, included between equidistant ordinates, the common interval between which is 1°, which we may take as the unit of x. The area of any section whose base is the unit of x will be approximately equal, numerically, to the middle ordinate of that section, and the approximation is closer, the smaller the adopted unit of x is. Thus the numbers g may be regarded as equidistant ordinates, corresponding to the abscissas 1.5°, 2.5°, &c., which are entered in column (4). We can construct a gamma curve (52) to represent these ordinates, just as if they were the coefficients in the expansion of some polynomial, and the curve thus obtained will be a close approximation to the true curve of probability.

First, to find the centre of gravity of the terms in column (3), regarded as the masses of material points, multiply each into its lever arm or distance from the place of amplitude zero, as given in column (4), and set the resulting moments in column (5). Their sum is 5.197, and dividing this by the sum of the masses, which is unity, we get 5.2 nearly as the lever arm of the centre of gravity. Subtracting this from the numbers in column (4), we have the abscissas x of the masses referred to their centre of gravity, and we enter them in column (6). They are the errors of the several observed q'ntities, referred to their arithmetical mean as a standard. The squared q. m. error ε^2 is found just as in the case of a common probability curve. Multiplying the square of each error by its probability, we set the products in column (7). Their sum is $\varepsilon^2 = 4.647$, this being unchanged when divided by the sum of the probabilities which is unity. Likewise the cube of the c. m. inequality is found by multiplying the cube of each error by its probability, setting the result in column (8), and taking their algebraic sum, which is $\zeta^3 = 8.490$. Then by (39), the constants in the gamma curve are

$$a = \frac{2\epsilon^2}{\zeta^3} = \frac{2\times 4.647}{8.490} = 1.095, \qquad b = \epsilon^2 = 4.647. \tag{64}$$

Hence $a^2b = 5.572$, and (52) gives $K = 1.0151$. The adopted unit of x being the common interval between successive ordinates in column (3), which interval is represented by dx in (52), we have $dx = 1$, and

$$\log y = \bar{1}.26081 + 4.572 \log (1 + .19652x) - .47555x. \tag{65}$$

Giving to x in this equation its values from column (6) of the table, we obtain the values of y which are set in column (9). Their sum is unity as it should be. They represent the general form of the given series of probabilities pretty closely, as shown by the differences $g - y$ in column (10). The numbers y may be fairly presumed to come much nearer to the true probabilities than the numbers g do.

As we have noticed, the numbers y here are taken to represent the areas lying between equidistant ordinates. The probability that an observed amplitude will be between 3° and 6° is therefore approximately

$$.183 + .200 + .171 = .554.$$

To find the probability that it will fall between any limits which are fractions of a degree, we can make an interpolation by the method which I gave in the *Smithsonian Report* of 1871, p. 309.

If however it is desired to evaluate rigorously the area of the gamma curve between given limits, it will be best to take the equation (25), where the origin is at the point in which the curve meets the X axis, and write

$$Y = \frac{a}{\Gamma(a^2b)} (ax)^{a^2b-1} e^{-ax}. \tag{66}$$

A known formula, obtained by integration by parts, is

$$\int v^{n-1} e^{-v} dv = -e^{-v}[v^{n-1} + (n-1)v^{n-2} + (n-1)(n-2)v^{n-3} + \ldots \\ \ldots + (n-1)(n-2)\ldots 2.1] + C, \tag{67}$$

where n is supposed to be an integer, otherwise the series will not thus terminate. We have then,

$$\int_v^\infty v^{n-1} e^{-v} dv = v^{n-1} e^{-v} \left\{ 1 + \frac{n-1}{v} + \frac{(n-1)(n-2)}{v^2} + \ldots + \frac{(n-1)\ldots 2.1}{v^{n-1}} \right\}. \tag{68}$$

But (66) gives

$$\int_x^\infty Y dx = \frac{1}{\Gamma(a^2b)} \int^\infty (ax)^{a^2b-1} e^{-ax} d(ax),$$

so that by (51) and (68) we have

$$n = a^2b, \qquad v = ax,$$

$$\int_x^\infty Y dx = \left(\frac{v}{n}\right)^{n-1} e^{n-v} \left(\frac{1 + \frac{n-1}{v} + \frac{(n-1)(n-2)}{v^2} + \ldots + \frac{(n-1)\ldots 2.1}{v^{n-1}}}{\sqrt{(2\pi n)}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \&c.\right)} \right) \tag{69}$$

When $v>n$, and n is somewhat large, it makes no difference for our purposes whether n is an integer or not, because the series in the numerator will be so convergent that some of its last terms may be neglected, and if this is true for the two nearest integers above and below n, it is also true for n, even though it be fractional. The series does not always converge rapidly, but its terms are easily computed, each from the one that precedes it. To insure accuracy, this part of the work should be carried to two more places of decimals than are required in the sum of the series. To integrate between the limits x_1 and x_2, we take the difference of two integrals from x_1 to ∞ and from x_2 to ∞.

But when $v<n$, or when n is small, we can use by preference another formula, also obtained by integration by parts,

$$\int v^{n-1} e^{-v} dv = v^n e^{-v} \left\{ \frac{1}{n} + \frac{v}{n(n+1)} + \frac{v^2}{n(n+1)(n+2)} + \&c. \right\} + C, \quad (70)$$

where n need not be a whole number. Taking this integral between the limits 0 and v, C disappears, and we get by (66) as before

$$\left.\begin{aligned} n &= a^2 b, \qquad v = ax, \\ \int_0^x Y dx &= \frac{v^n e^{-v}}{\Gamma(n+1)} \left\{ 1 + \frac{v}{n+1} + \frac{v^2}{(n+1)(n+2)} + \&c. \right\}, \end{aligned}\right\} \quad (71)$$

or, with the expression for $\Gamma(n+1) = n\Gamma(n)$ from (51),

$$\int_0^x Y dx = \left(\frac{v}{n}\right)^n e^{n-v} \left\{ \frac{1 + \frac{v}{n+1} + \frac{v^2}{(n+1)(n+2)} + \&c.}{\sqrt{(2\pi n)}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \&c.\right)} \right\}. \quad (72)$$

When either one of the two integrals (69) and (72) is known, the other is known also, because

$$\int_0^x Y dx + \int_x^\infty Y dx = 1. \quad (73)$$

Now in Table II. we have, by (64),

$$ab = 5.088, \qquad n = 5.572.$$

To find the probability that a single observed amplitude will fall below 3° for instance, the upper limit of integration is

$$x = 5.088 - 1.7 - 0.5 = 2.888, \qquad \therefore v = 3.162,$$

and with these values of n and v, (72) gives

$$\int_0^x Y dx = .1413.$$

For the probability that the amplitude will exceed 6°, the lower limit is

$$x = 5.088 + 0.3 + 0.5 = 5.888, \therefore v = 6.447,$$

and (69) gives

$$\int_x^\infty Y dx = .3111.$$

The series was carried only so far as the factors $n - 1$, $n - 2$, &c, were positive, and as none of the terms were small enough to be neglected, it might be doubted whether the result is correct. But when (72) is used, with the same value of v, we get

$$\int_0^x Ydx = .6890,$$

and $.3111 + .6890 = 1$ nearly, as it should be, so that the sufficient accuracy of the other result is confirmed. Thus the probabilities that an amplitude will fall below 3°, or between 3° and 6°, or above 6° are as found by integration

.141, .548, .311,

and as found by addition of terms in column (9) of the table,

.138, .554, .308.

The differences existing are due to the fact that the terms y in the table are middle ordinates, while the integration gives areas. The area between two ordinates which are separated by a unit interval will be numerically a little greater or less than the ordinate at the middle of the interval, according as the curve there is convex or concave toward the X axis.

The representation of these observations by the computed gamma curve might have been made a little more accurate if the 309 observed amplitudes had been published and treated separately, instead of being grouped within intervals of 1° each. It is of course only an approximation to the truth when we take the middle of such an interval as the point whose position represents that of all the observations in the group, for the purpose of finding the centre of gravity of the whole series, and the deviations from it by which we estimate the q. m. error ε and the c. m. inequality ζ, and thence get the values of a and b. When the observations are separately given, ε^2 is found just as in constructing a common probability curve, and ζ^3 in like manner, only taking the cubes of the + and — errors instead of their squares. The unit of x may be chosen at pleasure.

We might have made small corrections in a and b on account of the fact that the errors x in our table are residuals and not true errors. The calculation, I think, would be as follows. Any particular true error is the algebraic sum of the residual error and the error of the mean from which the residuals are reckoned. The residual error and the error of the mean may be treated as approximately independent of each other. Denote by ε and ζ the q. m. error and c. m. inequality for a system of true errors. The (q.m. e.)^2 and (c. m. i.)^3 for the residuals are

$$\frac{[gx^2]}{[g]}, \quad \text{and} \quad \frac{[gx^3]}{[g]},$$

where [ ] signifies summation throughout the series. The q. m. error of the mean is nearly $\varepsilon \div \sqrt{m}$, where m denotes 309, the whole number of observations. We have then by (62)

$$\varepsilon^2 = \frac{[gx^2]}{[g]} + \frac{\varepsilon^2}{m}. \tag{74}$$

The approximate c. m. inequality for the mean is $\zeta \div m^{\frac{2}{3}}$ according to (60), and (63) gives

$$\zeta^3 = \frac{[gx^3]}{[g]} + \frac{\zeta^3}{m^2}. \tag{75}$$

From the above we get, since $[g] = 1$,

$$\varepsilon^2 = \left(\frac{m}{m-1}\right)[gx^2], \qquad \zeta^3 = \left(\frac{m^2}{m^2-1}\right)[gx^3]. \tag{76}$$

Now $[gx^2]$ and $[gx^3]$ are the two sums 4.647 and 8.490 of the numbers in columns (7) and (8) of our table, so that for the system of true errors we have

$$\varepsilon^2 = \frac{309}{308} \times 4.647 = 4.662, \qquad \zeta^3 = \frac{95481}{95480} \times 8.490 = 8.490,$$

and the corrected values of a and b are by (64)

$$a = 1.098, \qquad b = 4.662. \tag{77}$$

It will be noticed that while ε^2 is quite perceptibly larger for true errors than for residuals, ζ^3 is hardly increased at all. It seems reasonable that this should be so, for the residuals are reckoned from the place of the arithmetical mean as an origin, and the q. m. error is thereby made a minimum. Any change in the place of the origin must increase ε. But there is no such necessity in the case of ζ. A change of origin may increase its absolute value or may diminish it. According to our formula, the chances are that it will be very slightly increased.

If the observations were separately given, we should find, in like manner, first, that the square of the q. m. error is greater for true errors than for residuals, in the ratio of m to $m-1$; and secondly, that the absolute value of the cube of the c. m. inequality is greater also, in the ratio of m^2 to m^2-1. The first of these is a well known result.

In any given set of observations there will usually be some inequality on the + and — sides of the mean, even when the real law of error is symmetrical on both sides, so that the asymmetry is purely fortuitous. To decide whether ζ^3 as found from the residuals is fortuitous or not, we shall sometimes need to know what its probable value would be, on the assumption that the true errors are represented by x in the symmetrical curve

$$Y = \frac{h}{\sqrt{\pi}} e^{-h^2x^2}. \tag{78}$$

The whole number of possible errors, each taken a number of times proportional to the probability of its ocucrrence, is represented by

$$\int_{-\infty}^{\infty} Ydx = \frac{h}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-h^2x^2}dx = 1.$$

Hence the mean of the squares, as well as the sum of the squares, of the cubes of all the possible errors is represented by

$$\int_{-\infty}^{\infty} x^6\, Ydx = \frac{1}{h^6\sqrt{\pi}}\int_{-\infty}^{\infty} (hx)^6 e^{-h^2x^2}d(hx),$$

or, putting $hx = t$ and $h^2 = 1 \div 2\varepsilon^2$,

$$\int_{-\infty}^{\infty} x^6\, Ydx = \frac{8\varepsilon^6}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^6 e^{-t^2}\, dt = 15\varepsilon^6, \qquad (79)$$

the known value of the last definite integral being $\frac{15}{8}\sqrt{\pi}$. (Sturm, *Cours d' Analyse*, II. p. 19.) The probable value of the cube of a single error is found approximately by taking the square root of the result in (79) and multiplying it by .6745, which gives

$$\pm.6745\varepsilon^3\sqrt{15}.$$

The probable value of the mean of the cubes of m errors is therefore

$$(\zeta^3) = \pm.6745\varepsilon^3\sqrt{(15 \div m)}. \qquad (80)$$

This is a standard which the actual value of ζ^3 ought not very much to exceed, if the law of error is to be considered symmetrical.

For example, in the set of observations at p. 495 of Vol. II. of Chauvenet's *Astronomy*, $m = 40$ and $\varepsilon = .202$, and (80) gives

$$(\zeta^3) = \pm.00340.$$

Actually, the algebraic sum of the cubes of the residuals is –.1364, so that

$$\zeta^3 = \frac{-.1364}{40} = -.00341.$$

Of course such a very close agreement between the actual and the prob'le value would not often occur, but in this and other cases, where ζ^3 does not much exceed (ζ^3), we may infer that no real c. m. inequality exists, and that the true law of error is probably symmetrical as in (78).

On the other hand, for the set of observations in our Table II. we have $m = 309$ and $\varepsilon^2 = 4.662$, and (80) gives the probable value

$$(\zeta^3) = \pm 1.496.$$

The actual value is $\zeta^3 = 8.490$, being almost 6 times as great. The chances are something like 10000 to 1 against the fortuitous occurrence of an error 6 times as great as the probable error. We must infer that, as indeed a simple inspection of the observations in this case indicates, the c. m. ineq. here is not only apparent, but real; so that an unsymmetrical curve alone can represent the true law of error with reasonable accuracy.